METHOD AND SYSTEM FOR IMPROVING DATA QUALITY IN LARGE HYPERLINKED TEXT DATABASES USING PAGELETS AND TEMPLATES

Abstract of the Disclosure

A computing system and method clean a set of hypertext documents to minimize violations of a Hypertext Information Retrieval (IR) rule set. Then, the system and method performs an information retrieval operation on the resulting cleaned data. The cleaning process includes decomposing each page of the set of hypertext documents into one or more pagelets; identifying possible templates; and eliminating the templates from the data. Traditional IR search and mining algorithms can then be used to search on the remaining pagelets, as opposed to the original pages, to provide cleaner, more precise results.